Universidad Michoacana de San Nicolás de Hidalgo

Facultad de Ciencias Físico-Matemáticas
"Mat. Luis Manuel Rivera Gutiérrez"

"Métodos de Inferencia Estadística para la EMOVI"

T E S I S

Que para obtener el título de:
Licenciada en Ciencias Físico-Matemáticas

Presenta:

Andrea Bethsabe García Gutiérrez

Asesor:

Dr. José Elías Rodríguez Muñoz
Departamento de Matemáticas
Universidad de Guanajuato

M.C. Christian Morales Ontiveros
Licenciatura en Ciencias Físico-Matemáticas
Universidad Michoacana de San Nicolás de Hidalgo

Morelia, Michoacán Febrero 2019

Índice

Dedicatoria

A mis padres, por ser mis maestros de vida, un pilar esencial en mi formación académica y humana, por todo y por tanto.

A mi hermana por ser mi apoyo incondicional.

A todas las personas que hicieron posible este trabajo.

Agradecimientos

Al Dr. José Elías por haberme brindado la oportunidad de trabajar con él, por compartir sus conocimientos y su invaluable asesoría.

A la Universidad Michoacana de San Nicolás de Hidalgo por abrirme sus puertas y ofrecerme la dedicación y enseñanza de mis profesores que son parte fundamental de mi formación académica.

Para la elaboración de la presente tesis, la autora contó con el apoyo de la Fundación Espinosa Rugarcía (ESRU) y el Centro de Estudios Espinosa Yglesias (CEEY) a través del Programa de Becarios CEEY.

Resumen

El Centro de Estudios Espinosa Yglesias (CEEY) se dedicó a hacer una encuesta a nivel nacional acerca de la movilidad social, la Encuesta ESRU de Movilidad Social en México 2011 (EMOVI-2011). Se han realizado varios análisis de la encuesta EMOVI-2011 en términos de estimadores puntuales de índices de movilidad social respecto a educación, ocupación, riqueza y percepción. En este proyecto analizaremos la documentación técnica y base de datos que se recomienda incluir en un estudio por muestreo, propondremos métodos de inferencia que darán mayor soporte a las interpretaciones que surjan al analizar su información y mostraremos la aplicación de los métodos mencionados.

Abstract

The Centro de Estudios Espinosa Yglesias (CEEY) dedicated itself to make a nationwide survey about social mobility, the ESRU Survey of Social Mobility in Mexico 2011 (EMOVI-2011). Several analyzes of the EMOVI-2011 survey have been carried out in terms of specific estimators of social mobility indices regarding education, occupation, wealth and perception. In this project we will analyze the technical documentation and database that is recommended to be included in a study by sampling, we will propose methods of inference that will give greater support to the interpretations that arise when analyzing your information and we will show the application of the aforementioned methods.

Palabras clave: conglomerados últimos, intervalos de confianza, inferencia estadística, contraste de hipótesis, movilidad social.

Introducción

La movilidad social es el desplazamiento de las personas en los distintos niveles socioeconómicos, mientras haya más movilidad en el país querrá decir que existe la igualdad de oportunidades(von Mentz (2003)) independientemente de las condiciones de nacimiento, género o sexo. Por desgracia no se contaba con mucha información acerca de la movilidad en México, pero el Centro de Estudios Espinosa Yglesias (CEEY), que se ha preocupado por el tema en cuestión, ha realizado la Encuesta ESRU de Movilidad Social en México (EMOVI) ya en tres ocasiones EMOVI-2006, EMOVI-2011 y EMOVI-2017, ya que era una excelente herramienta para obtener los datos que han dado pie a analizar la movilidad social en México. En algunos de los análisis de la encuesta se realizaron estimaciones puntuales de índices para evaluar la movilidad social además de inferir tazas con dichos índices en áreas como educación, ocupación, riqueza y percepción.

Otra manera de detallar los resultados obtenidos del análisis de la información de la Encuesta ESRU de Movilidad Social en México 2011 (EMOVI-2011) es por inferencia estadística ya que dará mayor soporte científico a esta y a futuras encuestas que realice el CEEY y como consecuencia podría potenciar la aceptación de las políticas públicas a favor de la movilidad social. Para lograr lo mencionado anteriormente, en este trabajo proponemos estimadores de varianza de estimadores puntuales de índices de movilidad social, métodos de construcción de intervalos de confianza para los índices y métodos de contraste de hipótesis para realizar comparaciones regionales y entre grupos sociales.

La secuencia de la exposición del presente trabajo es como sigue, en el capítulo 1 detallamos la documentación técnica y base de datos que se recomienda incluir cuando se realiza una encuesta con el propósito de dar a conocer las características de cómo funciona, cómo está diseñado y con qué fin se lleva a cabo la encuesta para que las personas interesadas en trabajar con su información comprendan la estructura. Como ejemplo tenemos la Encuesta Nacional de Ingresos y Gastos de los Hogares 2016 (ENIGH-2016) del INEGI que cuenta con la documentación completa. Posteriormente, la documentación técnica y base de datos de la EMOVI-2011. En el capítulo 2 presentamos los métodos para la obtención de estimadores de varianza, intervalos de confianza y el contraste de hipótesis que proponemos. En el capítulo 3 mostramos los resultados de aplicar los métodos de inferencia estadística propuestos a la información obtenida de la EMOVI-2011 sobre la movilidad.

1. Documentación técnica y base de datos en estudios basados en encuestas

Para administrar e incrementar la confiabilidad de un estudio por muestreo es recomendable tomar en cuenta varios documentos metodológicos que incluyan los distintos requerimientos para garantizar que se tiene una investigación completa tales como la definición del tema a investigar, la formulación de hipótesis hasta que procedimientos se necesitan para avalar la calidad de la información recabada y base de datos de la encuesta. Parte de lo que se presenta aquí se puede encontrar en el capítulo 17 de Särndal et al. (1992) y Sección 2.1 de Heeringa et al. (2010).

1.1. Descripción general

Para un entendimiento profundo de la investigación y sus componentes, los documentos técnicos y base de datos que recomendamos incluir en un estudio por muestreo son los siguientes:

- Diseño conceptual
    - Motivación
    - Antecedentes
    - Objetivos
- Marco de muestreo
- Base de datos
- Cuestionario
- Plan de muestreo
    - Unidad de muestreo
    - Población objetivo
    - Diseño de muestreo
        - Tipo de muestreo
        - Tamaño de la muestra
        - Procedimiento de selección

    - Probabilidades de inclusión
    - Factores de expansión
- Capacitación
- Análisis de la información de muestra
- Conclusiones y retroalimentación
- Reporte técnico y ejecutivo

En seguida explicamos en que consiste cada uno de ellos.

El diseño conceptual consta de tres elementos; motivación, antecedentes y objetivos. La motivación es el porqué y para qué de ejecutar una encuesta. Los antecedentes se refieren a si en años anteriores se realizó alguna encuesta donde se hallan propuesto variables u objetivos similares a los de la presente, ya que podría servir de guía, dar paso a comparaciones desde un enfoque estadístico o descubrir cuanto se conoce sobre el tema. Y los objetivos, que se definen al inicio de la investigación determinando lo que se pretende con su realización y la información que se obtenga. Se dividen en objetivo principal (o general), que pretende enunciar de manera precisa y contundente la meta que se persigue con la investigación; y específicos, que indican los logros en cada etapa de la investigación. Es necesario que sean claros, coherentes y realistas ya que su cumplimiento es importante para la satisfacción de los usuarios que requieren la encuesta.

El marco de muestreo es la información que ubica y dimensiona a todos los elementos de la población que se quiere estudiar, puede consistir de censos de vivienda, conteos de población, etc. De ahí se enlista la población objetivo para después seleccionar elementos para conformar una muestra. En las bases de datos del marco de muestreo debe estar todo lo necesario para poder seleccionar a los elementos de la población en la muestra.

La base de datos es el conjunto de datos obtenidos en al momento de realizar la encuesta, es de forma organizada y los datos están relacionados entre sí. Cada base de datos contiene tablas, donde las columnas guardan parte de la información sobre cada pregunta del cuestionario que se quiere guardar en la tabla y cada fila conforma un registro, en este caso, cada persona entrevistada. El diseño debe permitir añadir datos en cualquier momento,

poder editarlos o corregirlos, que sean de fácil acceso para proceder a su consulta, permitir la clasificación de la información, proporcionar la posibilidad de extraer o exportar datos para su análisis. Para este tipo de base se espera que cada registro este identificado por una llave que permita seleccionar a los individuos con los que se requiera trabajar; además, incluir los factores de expansión que se han aplicado a cada uno de los individuos seleccionados esto para poder realizar inferencia. La base de datos es acompañada por diccionarios y catálogos de codificación para facilitar la edición y consulta a las personas que diseñan o desean trabajar con la base de datos.

El cuestionario está conformado por una serie de preguntas redactadas de forma coherente y permiten obtener información necesaria para la investigación. Éstas pueden ser abiertas, cerradas o semiabiertas, a veces se combinan varios tipos de preguntas con el fin de poder utilizar diferentes técnicas de análisis de datos. En ellas deben estar reflejados los objetivos, permitir que exista homogeneidad, que sean concretas y por temas afines. Además, que se puedan identificar con facilidad las respuestas ya que eso permitiría una codificación satisfactoria de la encuesta, es decir, facilitar grabar los datos en las computadoras y formar la base de datos para poder tratar la información. Es importante que su orden sea lógico y no afecté las respuestas.

El siguiente documento es el plan de muestreo donde se define la unidad de muestreo, la población objetivo y el diseño de muestreo. Éste último se refiere al tipo y al tamaño de muestra, al procedimiento de selección, a las probabilidades de inclusión y factores de expansión. La unidad de muestreo es la unidad mínima de observación o el conjunto de unidades mínimas de las que se obtendrá información de las variables. Son seleccionadas del marco de muestreo. Por ejemplo, en los diseños de una sola etapa, la unidad de muestreo y el elemento (unidad mínima de observación) coinciden; y si se tratan de un diseño de más de una etapa, las unidades de muestreo están conformadas por más de un elemento. La población objetivo es el conjunto de individuos con características similares que les interesa investigar. Esta población excluye a los elementos que son difíciles de entrevistar, que sale costoso entrevistar o que por la definición de los objetivos no encajan en la muestra.

A continuación, se describen los elementos que tiene el diseño de muestreo. Existen distintos tipos de muestreo, pero se pueden clasificar en dos: probabilístico y no probabilístico. En el muestreo probabilístico se conoce la probabilidad de seleccionar la muestra o al menos la probabilidad de seleccionar cada elemento de la población en la muestra. Dentro de este tipo de

muestreo existe el muestreo aleatorio simple, aleatorio estratificado, aleatorio sistemático, aleatorio por conglomerados. En contraposición en el llamado muestreo no probabilístico no se tiene conocimiento de las mencionadas probabilidades. Contiene a el muestreo por cuotas, de conveniencia, bola de nieve y discrecional.

El tamaño de la muestra es el número de individuos que componen una muestra para poder estimar parámetros. Para obtener el tamaño se debe tomar en cuenta el tamaño de población, la varianza, el nivel de confianza, el error de estimación y su probabilidad asociada y el parámetro que interesa estimar. No hay un tamaño ideal que se pueda generalizar ya que varia dependiendo del tipo de muestreo, los objetivos y las características de la población, mientras más grande más costosa es y la precisión de las estimaciones se puede ver afectado.

Cuando se trata de muestra probabilística, el procedimiento de selección es un conjunto de experimentos (aleatorios) cuya probabilidad resultante de seleccionar una muestra coincide con la probabilidad del diseño de muestreo. Algunos tipos de procedimientos son: tómbola, números aleatorios o sistemática aleatoria. Para ejemplificar tenemos un muestreo aleatorio simple (MAS) sin reemplazo donde el conjunto formado por todas las muestras S tiene un total de muestras posibles:

$$C_{N,n} = \binom{N}{n},$$

luego la probabilidad de cada muestra esta dada por

$$P(S = s) = \frac{1}{\binom{N}{n}},$$

donde N es el tamaño de la población y n es el tamaño de la muestra.

Ahora, para seleccionar los elementos se puede utilizar el procedimiento por números aleatorios. En una hoja se generan números aleatorios entre 0 y 1, que son $e_1, e_2, \ldots, e_N$, que se asignan a los individuos; después se ordenan según el criterio elegido y se extraen n elementos que es el tamaño de la muestra.

Respecto a la probabilidad de inclusión, existen dos tipos: de primer orden, es la probabilidad de que cualquier individuo de la población este incluido en una muestra; y de segundo orden, es que cualquier par de individuos de la población tengan la probabilidad se estar incluidos en la muestra de manera simultánea. Los factores de expansión son la capacidad que tiene cada

individuo seleccionado de representar la población en la que está contenido, es el inverso de la probabilidad de inclusión.

Ahora, con referencia a la capacitación, cuando se ha determinado el cuestionario que se empleará en las entrevistas y después de elegir el diseño de muestreo se seleccionan los entrevistadores, que son un componente esencial en la recolección de información. Los criterios de selección de los entrevistadores son definidos por las personas encargadas de la investigación. Posteriormente, al ser seleccionadas las personas más competentes, se les proporciona una capacitación. La capacitación es un paso elemental porque la información debe ser de calidad, veraz y con las mínimas posibles alteraciones. Se les debe preparar para proceso de localización de los entrevistados, la realización de la entrevista, explicarles el funcionamiento del material que se les proporciona y estrategias para utilizar en el manejo de situaciones problemáticas.

El análisis de la información consiste en obtener un conocimiento detallado de cada una de las variables en la encuesta. Para ello se emplean diferentes métodos estadísticos, en particular los métodos de análisis descriptivo tales como tablas de frecuencia, representaciones gráficas como histogramas y medidas de resumen como promedios, porcentajes, varianzas, quintiles, etc. Depende del enfoque y del tipo de investigación que se haya seleccionado.

También se redactan los documentos de conclusión y retroalimentación que es la manera de dar a conocer los principales resultados y hallazgos a los que requirieron la investigación demandante de la misma y a la comunidad científica. Esto hará posible que se puedan comparar los resultados con otras investigaciones y dar a conocer dudas que quedan pendientes y nuevas interrogantes para investigaciones posteriores.

Y para finalizar recomendamos hacer un reporte técnico que es un documento que detalla la metodología utilizada en el diseño, la elaboración de los instrumentos adicionales que se utilizaron al momento de levantar la encuesta y los resultados de la investigación de forma concisa; en pocas palabras, el conjunto de documentos técnicos y base de datos que se revisaron en esta sección. También sugerimos redactar un reporte ejecutivo que es una presentación general breve que muestra los aspectos importantes de la investigación de forma clara y debe crear interés a quién lo está por el proyecto.

1.2. Ejemplo: ENIGH-2016

Desde hace varios años, el Instituto Nacional de Estadística, Geografía e Informática (INEGI) ha trabajado para la obtención de información estadística respecto a la población y su entorno en México. Han logrado estandarizar los procesos y metodologías por los cuales se consiguen los resultados. La documentación técnica y base de datos es esencial para un estudio por muestreo y el INEGI se ha preocupado para que sea muy completa. Para ejemplificarla utilizaremos la información (INEGI, 2017a) de la Encuesta Nacional de Ingresos y Gastos de los Hogares 2016 (ENIGH-2016).

Iniciamos la descripción de los documentos con el diseño conceptual que tiene como primera parte la motivación de la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) que es proporcionar información del comportamiento de los ingresos y gastos de los hogares.

Los antecedentes de la ENIGH-2016 son varios. Uno se realizó en 1956 y 1958 cuando la Dirección General de Estadística (DGE) que dependía de la Secretaría de Industria y Comercio (SIC) levantó la encuesta Ingresos y Egresos de la Población en México. En 1960 Las 16 Ciudades de la República Mexicana, Ingresos y Egresos Familiares. En 1963 y 1968, el Banco de México llevó a cabo el levantamiento de la encuesta Ingresos y Gastos Familiares. En el periodo de 1969-1970, nuevamente la DGE-SIC realizó la encuesta Ingresos y Egresos de la República Mexicana. En 1975, la Secretaría del Trabajo y Previsión Social (STPS) la efectuó. En 1977 la DGE, como parte de la Secretaría de Programación y Presupuesto (SPP), desarrolló la Encuesta Nacional de Ingreso-Gasto de los Hogares, trabajo que constituyó el antecedente inmediato de las encuestas que ha llevado a cabo la DGE-INEGI para los periodos: 1984, 1989, 1992, 1994, 1996, 1998, 2000, 2002, 2004, 2005, 2006, 2008, 2010, 2012, 2014.

Como objetivos la ENIGH-2016 tiene proporcionar un panorama estadístico del comportamiento de los ingresos y gastos de los hogares en cuanto a su monto, procedencia y distribución, ofrecer información sobre las características ocupacionales y sociodemográficas de los integrantes del hogar y las características de la infraestructura de la vivienda y el equipamiento del hogar.

El marco de muestreo está conformado por la información demográfica y cartográfica del Censo de Población y Vivienda 2010.

La base de datos de la ENIGH-2016 está conformada por 11 tablas de datos normalizadas en las que se distribuye la información obtenida de la

encuesta de acuerdo con los temas de mayor interés y con los objetivos de la encuesta para realizar análisis y tabulados. Para que su consulta sea fluida y accesible se incluye un diccionario de datos y un manual de criterios de validación. Es importante mencionar que esta base datos contiene los factores de expansión de primer orden para cada individuo.

El levantamiento de la encuesta se realizó del 21 de agosto al 28 de noviembre de 2016, la recolección de la información se hizo con distintos cuestionarios. A partir del año 2008, se llegó a un acuerdo de realizar ajustes a diversas preguntas en los diferentes cuestionarios esto con la finalidad de adaptarlos a la realidad de los hogares. El entrevistador tenía 6 posibles cuestionarios para la entrevista:

- Cuestionario de hogares y vivienda, jefe(a) del hogar.
- Cuestionario para personas de 12 o más años, informante directo.
- Cuestionario para negocios del hogar, responsable del negocio.
- Cuestionario para personas menores de 12 años, el encargado del menor.
- Cuadernillo de gastos diarios, encargado de realizar el gasto en alimentos y bebidas.
- Cuestionario de gastos del hogar, integrantes del hogar que realicen gastos.

Con relación al plan de muestreo lo que la ENIGH-2016 proporcionó es lo subsiguiente. El diseño de muestreo es de tipo por probabilístico; además, es bietápico, estratificado y por conglomerados. La unidad de muestreo es el hogar y la población objetivo son los hogares de nacionales o extranjeros, que residen habitualmente en viviendas particulares dentro del territorio nacional. El tamaño de la muestra fue de 81 515 viviendas.

Para la selección de la muestra se utilizó el Marco Nacional de Viviendas 2012 del INEGI, construido a partir del Censo de Población y Vivienda 2010. La denominan la muestra maestra, su diseño es probabilístico, estratificado, unietápico y por conglomerados. Está conformada por unidades primarias de muestreo (UPM) de todo México y son agrupaciones de viviendas con distintas características dependiendo de la clasificación: urbano alto, complemento urbano y rural. Posteriormente se realiza una estratificación con criterios determinados por el INEGI para que cada UPM sea parte de un único estrato geográfico y uno sociodemográfico. El total de estratos en el territorio nacional es de 683. Las UPM de la muestra maestra fueron seleccionadas por medio de un muestreo con probabilidad proporcional al tamaño.

La probabilidad de muestreo y los factores de expansión se obtuvieron para tres ámbitos: urbano alto, complemento urbano y rural. Los muestra INEGI (2017a) en el documento *Diseño muestral*.

En cuanto al reclutamiento de los entrevistadores se realizó entre el personal que había laborado en algún otro proyecto del INEGI. La capacitación se realizó en cuatro etapas. En la primera etapa se llevó a cabo la capacitación de los Responsables Estatales del Proyecto, Coordinadores Instructores Supervisores Estatales, Instructores Supervisores Estatales y Supervisores Regionales; en la segunda etapa se realizó la capacitación para los Entrevistadores, Supervisores, Jefe de Supervisores, Responsables de Captura Validación y Analistas, en las propias oficinas. En la tercera etapa se impartió la capacitación para los responsables de la Captura Validación y los Analistas; en la cuarta etapa se dio la capacitación a los Capturistas.

El INEGI no tiene como objetivo analizar la información que obtienen de la encuesta. Sin embargo, realizó un análisis estadístico de carácter descriptivo, se hizo una evaluación de los errores de muestreo de las principales estimaciones. Se han presentado varios documentos que muestran los resultados de la ENIGH-2016, pero los más destacados son que se estima un crecimiento en el ingreso promedio trimestral por hogar no visto anteriormente y que las variaciones referentes al ingreso son alrededor del 11 % a nivel nacional por deciles de hogares.

Como principales hallazgos se tienen que, en comparación con ediciones anteriores, la ENIGH-2016 contó con un tamaño de muestra de más de 81 mil viviendas, el más grande en la historia del país para una encuesta de ingresos y gastos. Con este tamaño de muestra, es posible generar estimaciones por entidad federativa con desagregación urbana (localidades de 2 500 habitantes y más) y rural (localidades con menos de 2 500 habitantes). Se midió tanto el ingreso como el gasto en todos los hogares encuestados.

Para finalizar se esperaría ver las conclusiones, un reporte técnico y ejecutivo de la encuesta. El reporte técnico se ha descrito en distintos documentos que proporcionan y que revisamos a lo largo de esta sección. Las conclusiones y el reporte ejecutivo se encuentran en INEGI (2017b) donde se destaca que las encuestas de ingresos y gastos han evolucionado para incluir el uso de cuestionarios estandarizados, modelos probabilísticos y levantamientos que cumplen con una regularidad cíclica. Esta evolución ha permitido que las encuestas se conviertan en insumos importantes para la definición de política pública y la medición de la eficacia de la misma en México. Principalmente la ENIGH-2016 permite, por primera vez en la historia del país, identificar

los ingresos y gastos de los hogares a nivel estatal, y dentro de cada entidad con corte urbano y rural. Los resultados parecen mostrar que las variables relacionadas con las carencias sociales empleadas por el CONEVAL tienen una consistencia con la serie histórica.

La siguiente tabla muestra los documentos que es recomendable tener en un estudio por muestreo y los que la ENIGH-2016 incluye:

Documento	ENIGH-2016	Observaciones
Diseño conceptual		
Motivación	✓	
Antecedentes	✓	
Objetivos	✓	
Marco de muestreo	✓	No es accesible la base de datos.
Base de datos	✓	Incluyen factores de expansión.
Cuestionario	✓	
Plan de muestreo		
Unidad de muestreo	✓	
Población objetivo	✓	
Diseño de muestreo		
Tipo de muestreo	✓	
Tamaño de la muestra	✓	
Procedimiento de selección	✓	
Probabilidades de inclusión	✓	
Factores de expansión	✓	
Capacitación	✓	
Análisis de la información de muestra	✓	El análisis estadístico es descriptivo.
Conclusiones y retroalimentación	✓	
Reporte técnico y ejecutivo	✓	

1.3. EMOVI-2011

La Fundación ESRU preocupados por la movilidad social en México y por la falta de información acerca de ella, ha trabajado por medio del Centro de Estudios Espinosa Yglesias (CEEY) para recolectar y analizar esa información. Durante varios años han levantado encuestas respecto a este tema, en ésta sección nos interesamos en mostrar específicamente la información que la Encuesta ESRU de Movilidad Social en México 2011 (EMOVI-2011) tiene en su documentación técnica y base de datos.

Como primer elemento se tiene la motivación de la EMOVI-2011 que fue la necesidad de complementar la ya antes realizada en el 2006, es decir, obtener información de subgrupos poblacionales en México que no se habían tomado en cuenta. Para saber cuánto se conoce ya sobre el tema y realizar una comparación desde un enfoque estadístico se presenta el origen del principal antecedente de la EMOVI-2011 que es la Encuesta ESRU de Movilidad Social en México fueron en 2006.

El objetivo principal se divide en analizar el grado de movilidad social intergeneracional a nivel nacional y conocer la proporción en que los recursos económicos influyen para lograr oportunidades y bienestar socioeconómico y por ende el cómo son transmitidos por parte de los padres a los hijos. Los objetivos específicos son: estimar la probabilidad de que cualquier persona tenga las herramientas para definir, desarrollar y alcanzar proyectos laborales y de vida; y medir el logro a través del tiempo, de satisfactores objetivos (ingresos) y subjetivos (bienestar), del esfuerzo personal.

El marco de muestreo está conformado por información del Censo de Población y Vivienda 2010 y el II Conteo de Población y Vivienda 2005. Al momento de crear las bases de datos se verificó que la información sea correcta y se detectaron errores que se podían corregir, la consistencia de la información. El CEEY proporcionó la información que se dividió en 5 módulos:

1. El entrevistado y sus padres
2. La composición del hogar del entrevistado
3. Hermanos del entrevistado
4. Cónyuge del entrevistado
5. Hijos del entrevistado

Los datos se organizaron por casos (filas) y variables (columnas), los casos representan a cada una de las entrevistas completas y las variables corresponden a las preguntas contenidas en el cuestionario, incluidas las asociadas al proceso de selección de los entrevistados y de ubicación de sus viviendas, en conjunto sumó 2,521 variables. Además, un diccionario de datos también dividido por módulos y una guía de usuario que facilita la consulta de las bases.

El cuestionario fue diseñado por el Cómite Permanente del CEEY y Asesor del Programa de Movilidad Social junto con la empresa Investigaciones Sociales, Políticas y de Opinión Pública (INVESPOP). El cuestionario está confirmado por 185 preguntas de las cuales 19 son abiertas y 10 son semiabiertas.

El plan de muestreo contiene a la unidad de muestreo que es la vivienda y como población objetivo tendría las características de ser mujeres o hombres de entre 25 y 64 años, por tipo de jefatura del hogar y en ámbito urbano o no urbano.

Respecto al diseño de muestreo, la muestra fue probabilística y para asegurar que toda zona geográfica fuera tomada en cuenta se empleó la regionalización que tiene el INEGI, siete regiones socioeconómicas a nivel estatal (RSE) y la clasificación de 4 categorías del CONEVAL a nivel municipal (IRS); además, el esquema de muestreo fue de conglomerados y por etapas. Se estratificó en urbano y no urbano y el tamaño de la muestra fue de 11,240 entrevistas y efectivas fueron 11,001. Los procedimiento de selección, las probabilidades de inclusión y los factores de expansión se mostrarán más adelante en el Capítulo 3.

En total fueron seleccionadas 281 unidades primarias con la siguiente distribución: 210 para el ámbito urbano y 71 para el no urbano.

RSE	IRS	Municipios	Loc. No Urbanas
1	1	10	2
1	2	5	2
1	3	8	3
1	4	7	4
2	1	4	2
2	2	9	2
2	3	6	3
2	4	15	4
3	1	-	2
3	2	5	2
3	3	13	3
3	4	16	4
4	1	2	2
4	2	6	2
4	3	10	3
4	4	20	4
5	1	1	2
5	2	-	2
5	3	5	3
5	4	22	4
6	1	-	2
6	2	-	2
6	3	10	3
6	4	20	4
7	3	1	2
7	4	15	3
	TOTAL	210	71

Los encuestadores recibieron capacitación para el correcto uso del cuestionario. Se les mostró la clasificación de éste, los criterios de selección de vivienda y si se daba el caso de haber más de una persona con el perfil requerido y como utilizar el material adicional.

Se han realizado distintos análisis con la información obtenida por la EMOVI-2011, los principales resultados y conclusiones se encuentran en el *Informe Movilidad Social en México 2013: Imagina tu futuro* (Grajales et al., 2013). El contenido del reporte técnico es lo que analizamos en esta sección. Es aconsejable incluir el reporte ejecutivo que resuma los aspectos importantes de la investigación ya que genera mayor interés al lector.

En la siguiente tabla se muestra la documentación técnica y base de datos que incluye la EMOVI-2011:

Documento	EMOVI-2011	Observaciones
Diseño conceptual		
Motivación	✓	
Antecedentes	✓	
Objetivos	✓	
Marco de muestreo	✓	No es accesible la base de datos.
Base de datos	✓	No incluyen factores de expansión.
Cuestionario	✓	
Plan de muestreo		
Unidad de muestreo	✓	
Población objetivo	✓	
Diseño de muestreo		
Tipo de muestreo	✓	
Tamaño de la muestra	✓	
Procedimiento de selección	✓	
Probabilidades de inclusión	✓	
Factores de expansión	✓	
Capacitación	✓	
Análisis de la información de muestra	✓	
Conclusiones y retroalimentación	✓	
Reporte técnico y ejecutivo	✓	No cuenta con reporte ejecutivo como tal.

Los hallazgos que localizamos durante la revisión de las bases de datos son que no incluían los factores de expansión de la primera etapa de muestreo y que basándonos en la base de datos de "El entrevistado y sus padres", encontramos que sólo se contaba con 161 Municipios y 43 Localidades No Urbanas como se muestra en la tabla siguiente:

RSE	IRS	Municipios	Loc. No Urbanas
1	1	6	2
1	2	4	2
1	3	5	2
1	4	4	3
2	1	2	2
2	2	7	2
2	3	8	1
2	4	12	1
3	1	-	1
3	2	6	1
3	3	14	1
3	4	8	-
4	1	36	-
4	2	5	3
4	3	11	-
4	4	15	-
5	1	1	1
5	2	-	-
5	3	3	6
5	4	15	2
6	1	-	2
6	2	-	2
6	3	6	5
6	4	16	4
7	3	1	-
7	4	9	-
	TOTAL	161	43

2. Métodos de inferencia estadística

Para analizar estadísticamente los índices de movilidad, proponemos métodos de estimación de varianza basado en "Conglomerados Últimos". Éstos a su vez servirán para la estimación de intervalos de confianza y métodos de contraste de hipótesis para comparaciones ya sea entre regiones o grupos sociales. Antes, mostramos los elementos necesarios para llegar a los estimadores citados.

2.1. Estimador de la varianza del estimador de Horvitz-Thompson del total

El estimador de Horvitz-Thompson (para más detalle ver Horvitz and Thompson (1952)) sirve para estimar parámetros poblacionales lineales, como el total o la media poblacionales. El estimador HT de un total está dado por

$$\hat{t}_{HT} = \sum_{k \in U} w_k y_k S_k, \tag{1}$$

donde U es la población objetivo, $w_k = \frac{1}{\pi_k}$ es el factor de expansión para el individuo $k \in U$, π_k es la probabilidad de inclusión de primer orden para el mismo individuo, y_k representa el valor de la variable de interés y S_k es variable aleatoria que indica si el k-ésimo individuo está o no en la muestra. Estos indicadores se pueden conceptualizar como un vector $S = (S_1, \ldots, S_N)$. Este vector se denomina vector muestra.

Es posible mostrar que la varianza del estimador del total de HT está dado por:

$$Var(\hat{t}_{HT}) = \sum_{k \in U} w_k y_k^2 + \sum_{j \in U} \sum_{k \neq j \in U} \pi_{jk} w_j w_k y_j y_k - t^2, \tag{2}$$

donde π_{jk} es la probabilidad de inclusión de segundo orden para los individuos $j, k \in U$.

Es posible mostrar también que un estimador insesgado de la varianza en la fórmula (2) está dado por:

$$V_1 = \sum_{k \in U} w_k (w_k - 1) y_k^2 S_k + \sum_{j \in U} \sum_{k \neq j \in U} \frac{w_j w_k}{\pi_{jk}} (\pi_{jk} - \pi_j \pi_k) y_j y_k S_j S_k, \tag{3}$$

siempre que $\pi_{jk} > 0$ para todos $j, k \in U$.

Se puede observar que el anterior estimador depende de las probabilidades de inclusión de segundo orden. Estas cantidades no se conocen para algunos diseños de muestreo utilizados en la práctica o sus respectivos valores no se almacenan en la base de datos de la muestra seleccionada. Si esta es la situación, no es posible utilizar el estimador V_1 y será necesario usar otra alternativa para estimar la varianza en la fórmula (2). Una alternativa es utilizar el método de "Conglomerados Últimos" que se muestra en las siguientes secciones. Este método simplifica la estimación de la varianza porque acumula los componentes de la varianza para las etapas múltiples en una fórmula de etapa única que requiere sólo el conocimiento de los estratos de la etapa primaria ya que no es necesario estimar la contribución a la varianza de las subsecuentes etapas dado que dicha contribución es menor a la contribución de la primera. Utiliza únicamente los factores de expansión de la primera etapa de muestreo y las estimaciones a este mismo nivel.

2.2. Muestreo Multinomial

Esta sección es un preámbulo a la presentación del estimador de la varianza en (2) por conglomerados últimos.

El diseño de muestreo Multinomial es tal que el vector muestra se distribuye como:

$$S \sim Multinomial(n, q_1 \ldots, q_N), \tag{4}$$

donde n es el tamaño de muestra conocido, y q_N son las proporciones poblacionales.

Para obtener una muestra por este diseño, se repite n veces de forma independiente un experimento

$$Multinomial(1, q_1, \ldots, q_N). \tag{5}$$

De esta forma se obtienen muestras con reemplazo y tales que

$$\sum_{k \in U} s_k = n, \tag{6}$$

con $s_k \in \{0, \ldots, n\}$ y $E(S_k) = nq_k$ para todo $k \in U$.

Un estimador insesgado del total para este diseño está dado por:

$$\hat{t}_{HH} = \frac{1}{n} \sum_{k \in U} \frac{y_k}{q_k} S_k. \tag{7}$$

A este estimador se le conoce como el estimador de Hansen-Hurwitz (Hansen and Hurwitz, 1943) donde la suma es sobre los elementos de la muestra seleccionada.

Es posible mostrar que la varianza del estimador en la expresión (7) está dada por:

$$Var(\hat{t}_{HH}) = \frac{1}{n} \sum_{k \in U} \frac{1}{q_k} \left(\frac{y_k}{q_k} - t \right)^2. \tag{8}$$

Un estimador insesgado de la anterior varianza está dado por:

$$\widehat{Var}(\hat{t}_{HH}) = \frac{1}{n(n-1)} \sum_{k \in U} \frac{1}{q_k} \left(\frac{y_k}{q_k} - \hat{t}_{HH} \right)^2 S_k. \tag{9}$$

Ahora, inspirados en el estimador de la expresión (9), se puede proponer un estimador de la varianza del estimador de Horvitz-Thompson como:

$$V_2 = \frac{n}{n-1} \sum_{k \in U} \left(w_k y_k - \frac{1}{n} \hat{t}_{HT} \right)^2 S_k. \tag{10}$$

A este estimador se le conoce como estimador de la varianza por conglomerados últimos (Para más detalle ver Hansen et al. (1953) secciones 6.1, 6.7 y Särndal et al. (1992) secciones 4.6, 11.2). Cabe mencionar que este último estimador tiende a sobreestimar la varianza de la expresión (2). Sin embargo, dicho estimador sólo utiliza las probabilidades de inclusión de primer orden, o los respectivos factores de expansión, a diferencia del estimador de la expresión (3) que utiliza además las de segundo orden. Adicionalmente, el cálculo del estimador anterior se puede hacer, relativamente fácil, en una hoja de cálculo. Por ejemplo, si te tiene en una columna los factores de expansión y en otra los valores de la variable de interés, se construye una nueva columna con la multiplicación de las dos anteriores, seguido se calcula la varianza de esta nueva columna y el resultado se multiplica por el tamaño de la muestra.

2.3. Muestreo en etapas múltiples

En el contexto del muestreo multietápico aparece el estimador de la varianza por conglomerados últimos. Por esto, se presenta en esta sección.

El muestreo en etapas múltiples es un término para denotar una familia de diseños de muestreo. Es una forma de restringir la selección aleatoria y suele utilizarse cuando la población es muy extensa y requiere de mucho tiempo para obtener la muestra. Consiste en primero seleccionar una muestra de UPM que sería la primera etapa, y posteriormente se selecciona una muestra de USM (unidad secundaria de muestreo), segunda etapa, dentro de cada UPM seleccionada en la primera etapa. Si el objeto de la investigación no se encuentra en las etapas anteriores, se pueden utilizar sucesivamente tantas etapas como se requieran hasta llegar a la muestra final y en cada una de ellas utilizar una técnica de muestreo distinta.

Cabe mencionar que en ocasiones el marco de muestreo no contiene la lista de las USM pero sí la correspondiente a las UPM. Si este es el caso y antes de seleccionar las USM en la muestra, primero se construye el marco de muestreo a detalle de USM pero únicamente para las UPM seleccionadas en la primera etapa.

Para este tipo de diseños de muestreo, conviene primero conceptuar a la población como la unión de N^I UPM:

$$U = U_1^I \cup \ldots \cup U_{N^I}^I. \tag{11}$$

El tamaño de la j-ésima UPM se denotará por N_j^I. Así el tamaño de la población estará dado por:

$$N = \sum_{j=1}^{N^I} N_j^I. \tag{12}$$

De lo anterior es posible verificar que el total se puede escribir como:

$$\begin{aligned} t &= \sum_{j=1}^{N^I} t_j^I \\ &= \sum_{j=1}^{N^I} \sum_{k \in U_j^I} y_{k|j}^{II}, \end{aligned} \tag{13}$$

donde

$$t_j^I = \sum_{k \in U_j^I} y_{k|j}^{II} \tag{14}$$

y $y_{k|j}^{II}$ denota el valor de la característica de interés para la k-ésima USM de la j-ésima UPM.

El correspondiente estimador de Horvitz-Thompson para este tipo de diseños de muestreo está dado por

$$\hat{t}_{HT} = \sum_{j=1}^{N^I} \hat{t}_{\pi j}^I \frac{S_j^I}{\pi_j^I}, \tag{15}$$

donde S_j^I es la j-ésima coordenada del vector muestra de la primera etapa, π_j^I es la j-ésima probabilidad de inclusión de primer orden de la primera etapa y

$$\hat{t}_{\pi j}^I = \sum_{k \in U_j^I} y_{k|j}^{II} \frac{S_{k|j}^{II}}{\pi_{k|j}^{II}}, \tag{16}$$

es el estimador de Horvitz-Thompson del total de la j-ésima UPM. Adicionalmente, $S_{k|j}^{II}$ es la k-ésima coordenada del vector muestra de la segunda etapa en la j-ésima UPM, $\pi_{k|j}^{II}$ es la k-ésima probabilidad de inclusión de primer orden de la segunda etapa en la j-ésima UPM.

En este mismo orden de ideas, si se utilizan sólo dos etapas, es posible mostrar que la varianza del estimador en (15) está dada por:

$$Var(\hat{t}_{HT}) = V_{UPM} + V_{USM}, \tag{17}$$

donde V_{UPM} denota la contribución a la varianza por las UPM y está dado por:

$$V_{UPM} = \sum_{j=1} N^I \frac{(t_j^I)^2}{\pi_j^I} + \sum_{i=1}^{N^I} \sum_{i \neq j=1}^{N^I} \frac{\pi_{ij}^I}{\pi_i^I \pi_j^I} t_i^I t_j^I - t^2, \tag{18}$$

π_{ij}^I es la probabilidad de inclusión de segundo orden para las UPM i y j. Adicionalmente, V_{USM} es la contribución a la varianza por las USM y está dado por:

$$V_{USM} = \sum_{j=1}^{N^I} \frac{Var(\hat{t}_{\pi j}^I)}{\pi_j^I}. \tag{19}$$

Por otro lado, es posible mostrar que el estimador de la varianza (17) por conglomerados últimos está dado por:

$$V_3 = \frac{n^I}{n^I - 1} \sum_{j=1}^{N^I} \left(\frac{\hat{t}_{\pi j}^I}{\pi_j^I} - \frac{1}{n^I} \hat{t}_{HT} \right)^2 S_j^I, \tag{20}$$

donde n^I es el tamaño de la muestra de UPM.

Lo anterior se puede extender cuando la selección de la muestra es para más de dos etapas.

Es conveniente observar que:

I. Es más apropiado referirse a la fórmula (20) como estimador por conglomerados últimos en lugar del estimador de la expresión (10).

II. Si la unidad $l \in U$corresponde a la k-ésima USM de la j-ésima UPM, entonces su probabilidad de inclusión de primer orden resultante es:

$$\pi_l = \pi_{k|j}^{II} \times \pi_j^I, \tag{21}$$

el producto de la probabilidad de inclusión de primer orden de la primera etapa por la correspondiente de la segunda etapa.

III. Para aplicar la fórmula (20) en la estimación de la varianza (17), es necesario tener acceso a las probabilidades de inclusión de primer orden, o los respectivos factores de expansión, de cada una de las etapas de muestreo. Si el muestreo es de etapas múltiples, es suficiente con tener acceso a dichas probabilidades de la primera etapa y a la información suficiente para estimar totales para las UPM en la muestra.

2.4. Estimador de razón

Algunos de los estimadores de la EMOVI-2011 son estimadores de razón, por esto, se expone la presente sección.

El estimador de razón de dos totales está dado por:

$$\hat{R} = \frac{\hat{t}_{yHT}}{\hat{t}_{xHT}}. \tag{22}$$

Una aproximación de la varianza por el método de linealización de Taylor es:

$$Var(\hat{R}) \approx \sum_{k\in U} \frac{u_k^2}{\pi_k} + \sum_{j\in U} \sum_{j\neq k\in U} \frac{\pi_{jk}}{\pi_j \pi_k} u_j u_k - t_u^2, \tag{23}$$

donde $u_k = \frac{1}{t_x}(y_k - Rx_k)$ y $t_u = \sum_{k\in U} u_k$.

El correspondiente estimador de la varianza por conglomerados últimos está dado por:

$$\widehat{Var}(\hat{R}) = \frac{n}{n-1} \sum_{k\in U} \left(w_k \hat{u}_k - \frac{1}{n} \hat{t}_u \right)^2 S_k, \tag{24}$$

donde $\hat{u}_k = \frac{1}{\hat{t}_{xHT}}(y_k - \hat{R}x_k)$ y $\hat{t}_u = \sum_{k\in U} w_k \hat{u}_k S_k$.

Si la población estuviera dividida en H estratos, entonces el estimador de razón estará dado por:

$$\hat{R} = \frac{\hat{t}_{yHT}}{\hat{t}_{xHT}} = \frac{\sum_{h=1}^{H} \hat{t}_{yh}}{\sum_{h=1}^{H} \hat{t}_{xh}}, \tag{25}$$

donde $\hat{t}_{yh} = \sum_{k\in U_h} w_k y_k$ y $\hat{t}_{xh}$ está dado de forma similar.

El estimador de la varianza utilizando el método de linealización de Taylor(Särndal et al. (1992) sección 5.5.) combinado con el de conglomerados últimos está dado por:

$$\widehat{Var}(\hat{R}) = \sum_{h=1}^{H} \left\{ \frac{n_h}{n_h - 1} \sum_{k \in U_h} \left(w_k \hat{u}_k - \frac{1}{n_h} \hat{t}_{uh} \right)^2 S_k \right\}. \tag{26}$$

En la anterior fórmula si $k \in U_h$, entonces $\hat{u}_k = \frac{1}{\hat{t}_{xHT}}(y_k - \hat{R}x_k)$ y $\hat{t}_{uh} = \sum_{k \in U_h} w_k \hat{u}_k S_k$.

2.5. Estimación en dominios

Nuevamente, cuando en la EMOVI-2011 se distinguen los parámetros de interés por grupos sociales o regiones, sus respectivos estimadores corresponden a estimadores en dominios. Aquí los dominios son, precisamente, los grupos sociales o regiones citadas.

Un dominio D es una partición de la población U para la que se diseña una muestra independiente y la muestra de la población corresponde al conjunto de las muestras de los dominios

$$U = \bigcup_{d=1}^{D} U_d. \tag{27}$$

Cabe destacar que cuando seleccionamos una muestra de la población U, se selecciona sin tomar en cuenta su división en dominios pues se estaría en el caso de muestreo estratificado. Es importante mencionar que se selecciona así la muestra porque no se tiene información en el marco muestral de la asociación entre individuos y dominios.

Por otro lado, es conveniente construir a partir de la variable de interés dos variables que nos servirán para estimar totales en cada dominio (Para más detalle ver capítulo 10 de Särndal et al. (1992)). Estas variables son:

$$\begin{aligned} z_{dk} &= \begin{cases} 1 & si \quad k \in U_d \\ 0 & \text{en otro caso.} \end{cases} \\ y_{dk} = z_{dk} y_k &= \begin{cases} y_k & si \quad k \in U_d \\ 0 & \text{en otro caso.} \end{cases} \end{aligned} \tag{28}$$

Así, el tamaño del dominio d es $N_d = \sum_{k \in U} z_{dk}$ y el total de la variable de interés en el mismo dominio es $t_{yd} = \sum_{k \in U} y_{dk}$. Los respectivos estimadores de HT de los anteriores parámetros son respectivamente:

$$\begin{aligned} \hat{N}_d &= \sum_{k \in U} w_k z_{dk} S_k \\ \hat{t}_{yd} &= \sum_{k \in U} w_k y_{dk} S_k \end{aligned} \tag{29}$$

Ahora supongamos que nos interesa el promedio de una variable de interés, pero en dos dominios:

$$\begin{aligned} \mu_{y1} &= \frac{1}{N_1} t_{y1} \\ \mu_{y2} &= \frac{1}{N_2} t_{y2} \end{aligned} \tag{30}$$

con estimadores:

$$\begin{aligned} \hat{\mu}_{y1} &= \frac{1}{\hat{N}_1} \hat{t}_{y1} \\ \hat{\mu}_{y2} &= \frac{1}{\hat{N}_2} \hat{t}_{y2}. \end{aligned} \tag{31}$$

Es posible mostrar que el estimador de la varianza de la diferencia de estimadores $\hat{\mu}_{y1} - \hat{\mu}_{y2}$, el estimador de la diferencia de promedios entre dominios, combinando el método de linealización de Taylor y de conglomerados últimos, está dado por:

$$\widehat{Var}(\hat{\mu}_{y1} - \hat{\mu}_{y2}) = \sum_{h=1}^{H} \left\{ \frac{n_h}{n_h - 1} \sum_{k \in U_h} \left(w_k \hat{u}_k - \frac{1}{n_h} \hat{t}_{uh} \right)^2 S_k \right\}. \tag{32}$$

En la anterior fórmula si $k \in U_h$, entonces

$$\hat{u}_k = \frac{1}{\hat{N}_1}(y_{1k} - \hat{u}_{y1} z_{1k}) - \frac{1}{\hat{N}_2}(y_{2k} - \hat{u}_{y2} z_{2k})$$
$$\hat{t}_{uh} = \sum_{k \in U_h} w_k \hat{u}_k S_k.$$

El estimador de la expresión (32) se utilizará más adelante en el análisis estadístico de la EMOVI-2011.

2.6. Estimadores de varianza, intervalos de confianza y contraste de hipótesis

Los estimadores y métodos de inferencia presentados aquí se aplicarán en el siguiente capítulo para mostrar su utilidad en el análisis estadístico de la EMOVI-2011. Los métodos de inferencia que proponemos son estimación de varianza por conglomerados últimos, ya mostrado en las secciones anteriores, construcción de intervalos de confianza y contraste de hipótesis.

Uno de los estimadores de varianza que se propone es (20) para el estimador de la varianza de $\hat{t}_{HT}$. Si se trata de un estimador de razón se propone el método de la ecuación (26). Por otro lado, si se requiere trabajar con dominios, por ejemplo urbano y rural u hombres y mujeres, proponemos el estimador de varianza de la forma (32).

La estimación puntual de un parámetro es simplemente el valor del estadístico correspondiente, pero la probabilidad de que sea el valor correcto no es muy grande. Por eso proponemos la estimación de parámetros por medio de intervalos de confianza. Es un método que nos permite proporcionar sobre que valores podemos hallar el valor del parámetro que se está buscando con una alta probabilidad. El intervalo de confianza depende de la estimación del parámetro, la varianza estimada y el nivel de confianza que es el porcentaje $(1-\alpha)\,\%$ que es la probabilidad de encontrar el parámetro verdadero en el intervalo.

El intervalo de confianza que se propone para $\hat{t}_{HT}$ es:

$$IC_{1-\alpha}(\hat{t}_{HT}) = \left(\hat{t}_{HT} \pm z_{1-\alpha} \sqrt{\widehat{Var}(\hat{t}_{HT})} \right), \tag{33}$$

donde $z_{1-\alpha}$ es el cuantil de orden $(1-\alpha)$ de la distribución normal estándar.

Si se quiere inferir un estimador de razón se utiliza el estimador de varianza (26) para obtener el siguiente intervalo de confianza:

$$IC_{1-\alpha}(\hat{R}) = \left(\hat{R} \pm z_{1-\alpha} \sqrt{\widehat{Var}(\hat{R})} \right), \tag{34}$$

donde $z_{1-\alpha}$ es el cuantil de orden $(1-\alpha)$ de la distribución normal estándar.

Para estimación de dominios, en este caso dos, construimos el intervalo de $(1-\alpha)\,\%$ de confianza para la diferencia como:

$$\begin{aligned} IC_{1-\alpha}(\mu_{y1} - \mu_{y2}) = &\left[\hat{\mu}_{y1} - \hat{\mu}_{y2} - z_{1-\alpha} \sqrt{\widehat{Var}(\hat{\mu}_{y1} - \hat{\mu}_{y2})}, +\infty \right) \\ &\cap \left[dif_{min}, dif_{max} \right], \end{aligned} \tag{35}$$

donde dif_{min} y dif_{max} son la diferencia mínima y máxima espera, respectivamente de $\mu_{y1} - \mu_{y2}$ y $z_{1-\alpha}$ es el cuantil de orden $(1-\alpha)$ de la distribución normal estándar. Esta forma del intervalo está asociado al contraste de hipótesis expresado en (38).

El contraste de hipótesis es una prueba estadística que responde si una proposición respecto a un parámetro poblacional es aceptable o no. Se tienen dos hipótesis: la nula (H_0) y la alternativa (H_1). La hipótesis que estará bajo investigación es H_0, y dependiendo de los resultados que se obtengan se rechazará o no. Dos elementos importantes son el estadístico de prueba, que utilizaremos para tomar una decisión en el contraste es el intervalo de confianza que nos proporcionará un conjunto de valores donde puede encontrarse el parámetro; y la regla de decisión que es el criterio que se utiliza para decidir si aceptamos o no la hipótesis nula. El contraste de hipótesis puede ser bilateral (donde H_0 expresa una igualdad y, por consecuencia, H_1 expresa desigualdad) o unilateral (donde en al menos una hipótesis aparece un signo de mayor que o menor que).

Para el estimador $\hat{t}_{HT}$ se puede utilizar un contraste de hipótesis bilateral como

$$H_0 : t = t_0 \quad vs. \quad H_1 : t \neq t_0. \tag{36}$$

El estadístico de prueba es $IC_{1-\alpha}(t_{HT})$ de la expresión (33) y la regla de decisión es rechazar H_0 si $t_0 \notin IC_{1-\alpha}(t_{HT})$ con un nivel de confianza $1-\alpha$. Si se rechazará H_0, se puede interpretar como que la información en la muestra proporciona evidencia a favor de H_1.

Ahora, para el estimador de razón se propone un contraste bilateral también

$$H_0 : R = R_0 \quad vs. \quad H_1 : R \neq R_0. \tag{37}$$

El estadístico de prueba es $IC_{1-\alpha}(R)$ y la regla de decisión es rechazar H_0 si $R_0 \notin IC_{1-\alpha}(R)$ con un nivel de confianza $1 - \alpha$. Esto es, que si se llegara a rechazar H_0, se puede interpretar como que la información en la muestra proporciona evidencia a favor de H_1.

Cuando se quieren comparar parámetros en distintos dominios, por ejemplo con dos dominios, se puede utilizar el contraste de hipótesis unilateral siguiente:

$$\begin{array}{cc} H_0 : \mu_{y1} > \mu_{y2} & vs. \quad H_1 : \mu_{y1} \leq \mu_{y2} \\ (\mu_{y1} - \mu_{y2} > 0) & (\mu_{y1} - \mu_{y2} \leq 0) \end{array}. \tag{38}$$

Con el estadístico de prueba (35) y la regla de decisión es que se rechaza H_0 si $IC_{1-\alpha}(\mu_{y1} - \mu_{y2}) \ni 0$, es decir, si el intervalo de confianza contiene al cero; además, se puede traducir como que la información en la muestra proporciona evidencia a favor de H_1.

3. Aplicación a la EMOVI-2011

Los métodos de inferencia propuestos en la sección anterior se pueden aplicar a la EMOVI-2011 porque algunos de los parámetros de interés se estiman con estimadores de razón. Cabe aclarar que para la estimación de la varianza se propone utilizar (20) pero como no se cuenta con los factores de expansión se usará (10). Aplicamos los métodos para la inferencia de un parámetro, la comparación de dos parámetros y cuando dos parámetros se encuentran en diferentes dominios en el ámbito de percepción. Además, es importante entender el diseño de muestreo que fue usado para recolectar la información y se tiene a continuación.

3.1. Diseño de muestreo

Lo presentado aquí se extrajo de la documentación técnica descrita en la sección 1.3. Revisaremos que el diseño de muestreo de la EMOVI-2011 contenga el tipo de muestreo, el tamaño de la muestra, el procedimiento de selección, las probabilidades de inclusión y los factores de expansión.

Primero, encontramos que la muestra fue probabilística y para asegurar que toda zona geográfica fuera tomada en cuenta se empleó la regionalización que tiene el INEGI, siete regiones socioeconómicas a nivel estatal (RSE) y la clasificación de 4 categorías del CONEVAL a nivel municipal (IRS); además, el esquema de muestreo fue de conglomerados y por etapas. Se estratificó en urbano y no urbano.

El tamaño de la muestra resultó de 11,240 entrevistas y efectivas fueron 11,001. La fórmula para obtenerla fue:

$$n = \frac{Z^2_{\alpha/2} P(1-P)\mathit{deff}}{\delta^2 TR},$$

donde n es el tamaño de la muestra, P es la proporción a estimar, $Z^2_{\alpha/2}$ es el cuantil de la distribución normal asociado al 95 % de confianza, δ es el margen del error máximo absoluto, $P(|P - \hat{P}| \leq \delta) = 0.95$, TRes la tasa de respuesta esperada, *deff* es el efecto de diseño.

A continuación mostramos como se dividió el proceso al ser el esquema de muestreo por conglomerados y por etapas.

En el estrato urbano fueron cuatro etapas:

- Primera etapa: Se realizó muestreo sistemático con probabilidad proporcional al tamaño (PPT) para la selección de las unidades primarias de muestreo (UPM) en este caso municipios donde

$$n_i^I = 210$$

- Segunda etapa: Se seleccionaron unidades secundarias de muestreo (USM) que eran áreas geoestadísticas básicas, realizando muestreo sistemático con probabilidad proporcional al tamaño (PPT) donde

$$n_j^{II} = 5$$

- Tercera etapa: Se realizó muestro aleatorio simple sin reemplazo obteniendo manzanas como unidades donde

$$n_k^{III} = 2$$

- Cuarta etapa: La selección se hizo por muestreo por cuotas, donde la característica era elegir una vivienda por lado de la manzana, suponiendo que era poligonal

$$n_l^{IV} = 1$$

Las probabilidades de inclusión fueron:

* UPM:

$$\pi_{ij} = \frac{N_{ij}}{N_j} m_j$$

donde:
π_{ij} es la probabilidad de seleccionar en una extracción con reemplazo el Municipio i del estrato j
N_{ij} es el total de la población de 25 a 64 años del Municipio ij
N_j es el total de la población de 25 a 64 años del estrato j
m_j es el número de Municipios seleccionados en el estrato j

* USM (AGEB):

$$\pi_{ijk} = \frac{N_{ijk}}{N_j} 5$$

donde:
π_{ijk} es la probabilidad proporcional al tamaño de seleccionar el AGEB k en el Municipio ij
N_{ijk} es el total de la población objetivo en el AGEB ijk
N_{ij} es el total de la población de 25 a 64 años del Municipio ij

* Manzanas:

$$\pi_{ijkl} = \frac{2}{M}$$

donde:
π_{ijkl} es la probabilidad de seleccionar la manzana l en el AGEB ijk
M es el total de manzanas en el AGEB ijk

* Viviendas:

$$\pi_{ijklm} = \frac{4}{V}$$

donde:
π_{ijklm} es la probabilidad de seleccionar la vivienda m en la manzana $ijkl$
V es el total de viviendas en la manzana $ijkl$

Y el factor de expansión es:

$$w_{ijklm} = \frac{1}{\pi_{ij}\pi_{ijk}\pi_{ijkl}\pi_{ijklm}}$$

En el estrato no urbano fueron dos etapas:

- Primera etapa: Se realizó muestreo sistemático con probabilidad proporcional al tamaño (PPT) para la selección de las unidades primarias de muestreo (UPM) en este caso localidades no urbanas donde

$$n_j^I = 71$$

- Segunda etapa: Se realizó muestreo por cuotas con el requisito de que cada localidad no urbana se dividiera en cuadrantes seleccionando dos de ellos para recorrerlos en forma de espiral y levantar el total de entrevistas donde

$$n_h^{II} = 40$$

Las probabilidades de inclusión fueron:

⋆ UPM:

$$\pi_{jh} = \frac{N_{hj}}{N_j} n_j$$

donde:
π_{jh} es la probabilidad de seleccionar en una extracción con reemplazo la localidad h del estrato j
N_{hj} es el total de la población de 25 a 64 años de la localidad hj
N_j es el total de la población de 25 a 64 años del estrato j
n_j es el número de localidades seleccionadas en el estrato j

⋆ USM (AGEB):

$$\pi_{jhm} = \frac{40}{N_{hj}}$$

donde:
π_{jhm} es la probabilidad de seleccionar el AGEB m en la localidad h del estrato j

N_{hj} es el total de la población de 25 a 64 años de la localidad hj

Y el factor de expansión es:

$$w_{jhm} = \frac{1}{\pi_{jh}\pi_{jhm}}$$

El total de estratos es $H = 204$ como se muestra en los hallazgos en la sección 1.3.

El objetivo de las secciones siguientes es mostrar la aplicación a la EMOVI-2011 de los métodos de estimación e inferencia estadística expuestos anteriormente.

3.2. Inferencia estadística: causas de éxito y fracaso

Como primer ejemplo tenemos que una manera de identificar las causas de baja o alta movilidad es a través de la percepción que tienen los entrevistados. Se realizará la prueba de hipótesis en base a la respuesta de la preguntas "173.¿Qué factores cree usted que son las dos causas más frecuentes por las que las personas sean pobres?"tomando la primer causa y "174.¿Qué factores cree usted que son los dos más importantes para tener éxito económico en la vida?"tomando el primer factor. La causa principal que identifican para obtener éxito es la iniciativa personal que corresponde a la principal causa de pobreza que es la flojera.

Tenemos como estimadores $\hat{q}_{173\_1,2}$ [1] y $\hat{q}_{174\_1,1}$[2]. Para el análisis utilizaremos el estimador de razón ya que la población esta dividida por estratos.

Para la causa de fracaso tenemos

$$\hat{q}_{173\_1,2} = \frac{\hat{t}_f}{\hat{N}} = 0.3002,$$

donde $\hat{t}_f$ es el estimador del total de la variable de interés que indica si la respuesta a la pregunta 173 fue "La flojera y falta de iniciativa", $\hat{N}$ el estimador del tamaño. La estimación de la varianza para calcular el intervalo de confianza se obtuvo con la ecuación (26) dando como resultado

$$\widehat{Var}(\hat{q}_{173\_1,2}) = 0.0000949$$

Y el intervalo de confianza del 95 % correspondiente es

$$\begin{aligned} IC_{95\,\%}(q_{173_1,2}) &= \left(\hat{q}_{173\_1,2} \pm z_{0.975}\sqrt{\widehat{Var}(\hat{q}_{173\_1,2})}\right) \\ &= (0.2811, 0.3193) \end{aligned}$$

Esto se puede interpretar como que la proporción población $q_{173_1,2}$ está entre 0.2811 y 0.3193 con probabilidad del 95 %.

[1] En la base de datos se tiene la variable correspondiente a la pregunta 173 como 173_1 donde la flojera es la opción número 2.

[2] En la base de datos se tiene la variable correspondiente a la pregunta 174 como 174_1 donde la flojera es la opción número 1.

Y para la causa de éxito obtuvimos que

$$\hat{q}_{174\_1,1} = \frac{\hat{t}_e}{\hat{N}} = 0.3020,$$

donde $\hat{t}_e$ es el estimador del total de la variable que indica si la respuesta a la pregunta 174 fue "Iniciativa personal", $\hat{N}$ el estimador del tamaño. La estimación de la varianza también se calculó con la ecuación (26) dando como resultado

$$\widehat{Var}(\hat{q}_{174\_1,1}) = 0.0000898$$

Y con un intervalo de confianza del 95 %

$$\begin{aligned} IC_{95\,\%}(q_{1734_1,1}) &= \left(\hat{q}_{174\_1,1} \pm z_{0.975}\sqrt{\widehat{Var}(\hat{q}_{174\_1,1})}\right) \\ &= (0.2834, 0.3206) \end{aligned}$$

De manera similar, lo anterior se puede interpretar como que la proporción de individuos que expresaron que la principal causa para obtener éxito es la iniciativa personal que está entre 0.2834 y 0.3206 con probabilidad del 95 %. El nivel de confianza expresa la probabilidad de acertar en la estimación, en este caso sera del 95 %.

Ahora, queremos saber si las proporciones de la principal causa de pobreza y la principal causa de éxito son iguales. Realizamos el contraste de hipótesis que consiste en que la hipótesis nula H_0 se refiere a que los parámetros $q_{173\_1,2}$ y $q_{174\_1,1}$ son iguales, y la hipótesis alternativa H_1 donde los parámetros son distintos, y se muestran a continuación:

$$H_0 : q_{173\_1,2} = q_{174\_1,1} \quad \text{vs.} \quad H_1 : q_{173\_1,2} \neq q_{174\_1,1}$$

El estadístico de prueba es $IC_{95\,\%}(q_{173\_1,2} - q_{174\_1,1})$ y la regla de decisión es:

- Rechazar H_0 si $0 \notin IC_{95\,\%}(q_{173\_1,2} - q_{174\_1,1})$, es decir, si rechazamos H_0 la información en la muestra proporciona evidencia a favor de la hipótesis H_1 con un nivel de confianza del 95 %.

Tenemos que

$$IC_{95\,\%}(q_{173\_1,2} - q_{174\_1,1}) = \left(\hat{q}_{173\_1,2} - \hat{q}_{174\_1,1} \pm z_{0.975}\sqrt{\widehat{Var}(\hat{q}_{173\_1,2} - \hat{q}_{174\_1,1})}\right)$$
$$= (-0.0378, 0.0342)$$

Dado que el anterior intervalo de confianza contiene al cero, se verifica la regla de decisión resultando que la H_0 no se rechaza. Se puede interpretar como que las proporciones de las causas antes citadas tienen una alta probabilidad de ser iguales; con base a la evidencia de la información de la muestra.

3.3. Inferencia estadística: movilidad mujeres y hombres

Otro análisis de la movilidad se basa en el género. En este caso se considera la estimación en dominios que serían dos: mujeres y hombres. Se crea una variable en base a las preguntas: "124. Comparando el hogar donde vivía a los 14 años, con todos los hogares de México en este tiempo, en una escala de 1 a 10, en la que 1 son los hogares más pobres y 10 son los más ricos, ¿dónde pondría usted su hogar de ese entonces?", la otra pregunta es "169.Comparando este hogar con todos los hogares de México en este momento, en una escala de 1 a 10, en la que 1 son los hogares más pobres y 10 son los más ricos, ¿dónde pondría usted este hogar?", la variable es movilidad que indica si existe un avance en el nivel socioeconómicos con base en su hogar de origen.

Como contraste de hipótesis se propone que

$$H_0 : m_m > m_h \quad \text{vs.} \quad H_1 : m_m \leq m_h$$

donde m_m el el promedio de la movilidad en mujeres y m_h el promedio de la movilidad en hombres ambos de la forma de la expresión (30). Donde H_0 expresa que en promedio las mujeres experimentan mayor movilidad que los hombres y H_1 que el promedio de la movilidad de los hombres es mayor o igual que el de las mujeres.

El estadístico de prueba es $IC_{95\,\%}(m_m - m_h)$ y la regla de decisión es:

- Rechazar H_0 si $0 \in IC_{95\,\%}(m_m - m_h)$, si el intervalo de confianza contiene al cero.

Tenemos que

$$\hat{m}_m = \frac{\hat{t}_{mm}}{\hat{N}_m} = 0.3073$$

donde $\hat{t}_{mm}$ es el estimador del total de la movilidad en las mujeres y $\hat{N}_m$ es el estimador del total de mujeres.

$$\hat{m}_h = \frac{\hat{t}_{mh}}{\hat{N}_h} = 0.2953$$

donde $\hat{t}_{mh}$ es el estimador del total de la movilidad en las hombres y $\hat{N}_h$ es el estimador del total de hombres.

La varianza la estimamos con la forma de la ecuación (32) dando como resultado

$$\widehat{Var}(\hat{m}_m - \hat{m}_h) = 0.00048$$

El intervalo de confianza con un 95 % es

$$\begin{aligned} IC_{95\,\%}(m_m - m_h) &= \left(\hat{m}_m - \hat{m}_h - z_{0.95}\sqrt{\widehat{Var}(\hat{m}_m - \hat{m}_h)}, +\infty\right) \cap [-1, 1] \\ &= (-0.0240, 1] \end{aligned}$$

Como el intervalo de confianza contiene al cero, se rechaza H_0.

Con base a la evidencia de ña información de la muestra, lo anterior se puede interpretar como que la percepción de movilidad promedio de los hombres es al menos igual a la percepción promedio de las mujeres. Las posibles causas de este comportamiento de la percepción de movilidad entre hombres y mujeres pueden ser de interés para los especialistas en este tema.

Conclusiones

Mostramos que para contribuir en la confiabilidad al realizar un muestreo se necesita revisar la documentación técnica para la encuesta y su base datos. Esto incluye la revisión de informes como documentos que incorporen el análisis de los datos para la comprensión de los datos. Debe incluir como mínimo una descripción resumida de la muestra una discusión sobre los procedimientos de ponderación y estimación. El INEGI es uno de los organismos que se preocupa por el proceso y tratamiento de la información y tienen completa su documentación. El CEEY también es muy cuidadoso con su documentación ya que es importante otorgar información completa a los investigadores que desean trabajar en conjunto. Lo anterior se describió en el capítulo 1.

Cuando revisamos los trabajos publicados respecto a la EMOVI-2006 y EMOVI-2011 pudimos identificar que se contaba con ciertos análisis de la información, esto dio oportunidad de complementar análisis con inferencia estadística. Propusimos métodos que se pueden utilizar con distintos parámetros dependiendo de la hipótesis que se desea contrastar. Y finalmente, dimos ejemplos de la aplicación de los métodos con los datos de movilidad perspectiva. Esto fue expuesto en los capítulos 2 y 3 respectivamente.

Es conveniente señalar los hallazgos de la investigación, la formación de algunos estratos no coinciden con la documentación del diseño de muestreo dada por el CEEY, esto apoyado en la información de las bases de datos que se proporcionaron. Seguido de que los factores de expansión no da una estimación adecuada de las viviendas de la población al momento de levantarse la encuesta. Adicionalmente, no se encontraron los factores de expansión de la primera etapa de muestreo, lo cual es necesario para estimar la varianza por el método de conglomerados últimos.

Finalmente, presentando estos métodos de inferencia estadística obtenidos esperamos brindar un mayor soporte científico a posibles interpretaciones y análisis en las distintas líneas de investigación y en otras áreas del conocimiento para esta y futuras encuestas EMOVI.

Anexos

Proceso de estimación de varianza desde la percepción de fracaso y éxito efectuado en el programa R Studio utilizando el estimador de razón.

```
#Importamos la librería "dplyr"
library(dplyr)

#Se abre la base de datos en la que se trabajará
EMOVIPercep <- read.csv("C:/Users/Beth/Desktop/R_EMOVI/
    Emovi2011Percep.csv", header=TRUE, sep=",")

#Se filtran los valores que se tratarán para tener un
    total de 10,933 observaciones
EMOVIPercep<-EMOVIPercep[!(EMOVIPercep$p173_1 %in% c("
    Ns (esp.)","Nc (esp.)")),]
EMOVIPercep<-EMOVIPercep[!(EMOVIPercep$p174_1 %in% c("
    Ns (esp.)","Nc (esp.)")),]

#Se inicializan nuevas variables
EMOVIPercep$fracaso<-0
EMOVIPercep$exito<-0
EMOVIPercep$wfracaso<-0
EMOVIPercep$wexito<-0
EMOVIPercep$u_kf<-0
EMOVIPercep$wu_kf<-0
EMOVIPercep$u_ke<-0
EMOVIPercep$wu_ke<-0
EMOVIPercep$u_2<-0

#Se asigna el valor 1 si se eligió como causa de
    fracaso la "La flojera y falta de iniciativa" y 0 en
     otro caso
for(i in 1:10933) {if(EMOVIPercep$p173_1[i]=="La
    flojera y falta de iniciativa") EMOVIPercep$fracaso[
    i]<-1 else EMOVIPercep$fracaso[i]<-0}
```

```
for(i in 1:10933) {EMOVIPercep$wfracaso[i]<-(
   EMOVIPercep$ponderel[i]*EMOVIPercep$fracaso[i]) }

#Se asigna el valor 1 si se eligió como causa de éxito
   la "Iniciativa personal" y 0 en otro caso
for(i in 1:10933) {if(EMOVIPercep$p174_1[i]=="
   Iniciativa_personal") EMOVIPercep$exito[i]<-1 else
   EMOVIPercep$exito[i]<-0}
for(i in 1:10933) {EMOVIPercep$wexito[i]<-(EMOVIPercep$
   ponderel[i]*EMOVIPercep$exito[i]) }

#Obteniendo el promedio de las variables de interés
sumaFracaso=sum(EMOVIPercep$wfracaso)
sumaExito=sum(EMOVIPercep$wexito)
sumN=sum(EMOVIPercep$ponderel)

R_e=(sumaExito)/sumN
R_f=(sumaFracaso)/sumN


#Calculando el valor estimado de u_k para causas de
   fracaso y éxito
for(i in 1:10933) {EMOVIPercep$u_kf[i]<-((EMOVIPercep$
   fracaso[i]-R_f)/sumN) }
for(i in 1:10933) {EMOVIPercep$u_ke[i]<-((EMOVIPercep$
   exito[i]-R_e)/sumN) }

for(i in 1:10933) {EMOVIPercep$wu_kf[i]<-(EMOVIPercep$u
   _kf[i]*EMOVIPercep$ponderel[i]) }
for(i in 1:10933) {EMOVIPercep$wu_ke[i]<-(EMOVIPercep$u
   _ke[i]*EMOVIPercep$ponderel[i]) }
```

```
#####################################################
#Estimación de la varianza para causa de fracaso#
#####################################################

#Se agrupan los valores por Municipios y Localidades no
    Urbanas
group_mpo<-group_by(EMOVIPercep, strat, cve_mpo)
tab_f<-summarise(group_mpo, sum_f=sum(wu_kf), n_f=n(),
   t_tf=(sum(wu_kf)/n()))

#Se iniciliza una nueva variable
group_mpo$wu_2<-0
for(j in 1:204){for(i in 1:10933){if(group_mpo$strat[i
   ]==tab_f$strat[j] & group_mpo$cve_mpo[i]==tab_f$cve_
   mpo[j]) group_mpo$wu_2[i]<-((group_mpo$wu_kf[i]-tab_
   f$t_tf[j])^2)}}

tab_f<-summarise(group_mpo, sum_f=sum(wu_kf), n_f=n(),
   t_tf=(sum(wu_kf)/n()), s_2=sum(wu_2))

#Se iniciliza una nueva variable
tab_f$n_n<-0
for(j in 1:204) {tab_f$n_n[j]<-((tab_f$n_f[j]*tab_f$s_
   2[j])/(tab_f$n_f[j]-1))}

#Se calcula el valor de la estimación de varianza
vvar_cf=sum(tab_f$n_n)
var_cf=sqrt(sum(tab_f$n_n))

#Se construye el intervalo de confianza con 95%
CI_IF=R_f-(1.96*var_cf)
CI_SF=R_f+(1.96*var_cf)
```

```
##################################################
#Estimación de la varianza para causa de éxito#
##################################################

#Se agrupan los valores por Municipios y Localidades no
    Urbanas
group_mpoe<-group_by(EMOVIPercep, strat, cve_mpo)
tab_e<-summarise(group_mpoe, sum_e=sum(wu_ke), n_e=n(),
    t_te=(sum(wu_ke)/n()))

#Se iniciliza una nueva variable
group_mpoe$wu_2<-0
for(j in 1:204){for(i in 1:10933){if(group_mpoe$strat[i
   ]==tab_e$strat[j] & group_mpoe$cve_mpo[i]==tab_e$cve
   _mpo[j]) group_mpoe$wu_2[i]<-((group_mpoe$wu_ke[i]-
   tab_e$t_te[j])^2)}}

tab_e<-summarise(group_mpoe, sum_e=sum(wu_ke), n_e=n(),
    t_te=(sum(wu_ke)/n()), s_2=sum(wu_2))

#Se iniciliza una nueva variable
tab_e$n_n<-0
for(j in 1:204) {tab_e$n_n[j]<-((tab_e$n_e[j]*tab_e$s_
   2[j])/(tab_e$n_e[j]-1))}

#Se calcula el valor de la estimación de varianza
vvar_ce=sum(tab_e$n_n)
var_ce=sqrt(sum(tab_e$n_n))

#Se construye el intervalo de confianza con 95%
CI_IE=R_e-(1.96*var_ce)
CI_SE=R_e+(1.96*var_ce)
```

Proceso para aplicar los métodos de inferencia estadística como la estimación de varianza de la diferencia de las proporciones de fracaso y éxito efectuado en el programa R Studio.

```
#Importamos la librería "dplyr"
library(dplyr)

#Se abre la base de datos en la que se trabajará
EMOVIPercAmbos <- read.csv("C:/Users/Beth/Desktop/R_
    EMOVI/Emovi2011Percep.csv", header=TRUE, sep=",")

#Se filtran los valores que se tratarán para tener un
    total de 10,933 observaciones
EMOVIPercAmbos<-EMOVIPercAmbos[!(EMOVIPercAmbos$p173_1
    %in%c("Ns_(esp.)","Nc_(esp.)")),]
EMOVIPercAmbos<-EMOVIPercAmbos[!(EMOVIPercAmbos$p174_1
    %in%c("Ns_(esp.)","Nc_(esp.)")),]

#Se inicializan nuevas variables
EMOVIPercAmbos$fracaso<-0
EMOVIPercAmbos$exito<-0
EMOVIPercAmbos$wfracaso<-0
EMOVIPercAmbos$wexito<-0
EMOVIPercAmbos$u_k<-0
EMOVIPercAmbos$wu_k<-0




#Se asigna el valor 1 si se eligió como causa de
    fracaso la "La flojera y falta de iniciativa" y 0 en
     otro caso
for(i in 1:10933) {if(EMOVIPercAmbos$p173_1[i]=="La_
    flojera_y_falta_de_iniciativa") EMOVIPercAmbos$
    fracaso[i]<-1 else EMOVIPercAmbos$fracaso[i]<-0}
for(i in 1:10933) {EMOVIPercAmbos$wfracaso[i]<-(
    EMOVIPercAmbos$ponderel[i]*EMOVIPercAmbos$fracaso[i
    ]) }
```

```
#Se asigna el valor 1 si se eligió como causa de éxito
    la "Iniciativa personal" y 0 en otro caso
for(i in 1:10933) {if(EMOVIPercAmbos$p174_1[i]=="
    Iniciativa_personal") EMOVIPercAmbos$exito[i]<-1
    else EMOVIPercAmbos$exito[i]<-0}
for(i in 1:10933) {EMOVIPercAmbos$wexito[i]<-(
    EMOVIPercAmbos$ponderel[i]*EMOVIPercAmbos$exito[i])
    }

#Obteniendo el promedio de las variables de interés
sumaFracasoAmbos=sum(EMOVIPercAmbos$wfracaso)
sumaExitoAmbos=sum(EMOVIPercAmbos$wexito)
sumNAmbos=sum(EMOVIPercAmbos$ponderel)

q_174_1=(sumaExitoAmbos)/sumNAmbos
q_173_1=(sumaFracasoAmbos)/sumNAmbos


#Calculando el valor estimado de u_k
for(i in 1:10933) {EMOVIPercAmbos$u_k[i]<-((
    EMOVIPercAmbos$fracaso[i]-q_173_1)/sumNAmbos)-((
    EMOVIPercAmbos$exito[i]-q_174_1)/sumNAmbos) }
for(i in 1:10933) {EMOVIPercAmbos$wu_k[i]<-(
    EMOVIPercAmbos$u_k[i]*EMOVIPercAmbos$ponderel[i]) }


############################
#Estimación de la varianza#
############################

#Se agrupan los valores por Municipios y Localidades no
     Urbanas
group_ambos<-group_by(EMOVIPercAmbos, strat, cve_mpo)
tab_ambos<-summarise(group_ambos, sum_ambos=sum(wu_k),
    n_a=n(), t_ta=(sum(wu_k)/n()))

#Se iniciliza una nueva variable
```

```
group_ambos$wu_2<-0
for(j in 1:204){for(i in 1:10933){if(group_ambos$strat[
    i]==tab_ambos$strat[j] & group_ambos$cve_mpo[i]==tab
    _ambos$cve_mpo[j]) group_ambos$wu_2[i]<-((group_
    ambos$wu_k[i]-tab_ambos$t_ta[j])^2)}}

tab_ambos<-summarise(group_ambos, sum_ambos=sum(wu_k),
    n_a=n(), t_ta=(sum(wu_k)/n()), s_2=sum(wu_2))

#Se iniciliza una nueva variable
tab_ambos$n_n<-0
for(j in 1:204) {tab_ambos$n_n[j]<-((tab_ambos$n_a[j]*
    tab_ambos$s_2[j])/(tab_ambos$n_a[j]-1))}

#Se calcula el valor de la estimación de varianza
var_ambos=sqrt(sum(tab_ambos$n_n))

#Se construye el intervalo de confianza con 95%
CI_IA=(q_173_1-q_174_1)-(1.96*var_ambos)
CI_SA=(q_173_1-q_174_1)+(1.96*var_ambos)
```

Proceso para aplicar los métodos de inferencia estadística correspondientes a los datos acerca de movilidad de mujeres y hombres en el programa R Studio.

```
#Importamos la librería "dplyr"
library(dplyr)

#Se abre la base de datos en la que se trabajará
EMOVIPMov <- read.csv("C:/Users/Beth/Desktop/R_EMOVI/
    MOV/EMOVI2011PMov.csv", header=TRUE, sep=",")

#Se filtran los valores que se tratarán para tener un
    total de 10,864 observaciones
EMOVIPMov<-EMOVIPMov[!(EMOVIPMov$p124 %in% c("Ns (esp.)
    ","Nc (esp.)")),]
EMOVIPMov<-EMOVIPMov[!(EMOVIPMov$p169 %in% c("Ns","Nc")
    ),]

#Se inicializan nuevas variables
EMOVIPMov$ponderel_M<-0
EMOVIPMov$ponderel_H<-0
EMOVIPMov$movilidad<-0
EMOVIPMov$mov_M<-0
EMOVIPMov$mov_H<-0

for(i in 1:10864){ if(EMOVIPMov$sex_ent[i]=="Mujer")
    EMOVIPMov$ponderel_M[i]<-EMOVIPMov$ponderel[i] else
    EMOVIPMov$ponderel_M[i]<-0}
for(i in 1:10864){ if(EMOVIPMov$sex_ent[i]=="Hombre")
    EMOVIPMov$ponderel_H[i]<-EMOVIPMov$ponderel[i] else
    EMOVIPMov$ponderel_H[i]<-0}




#Cambio de puntajes de clasificación de 1-5 del hogar
```

```
    de origen
for(i in 1:10864){ if(EMOVIPMov$p124[i]==1 | EMOVIPMov$
    p124[i]==2) EMOVIPMov$p124[i]<-1}
for(i in 1:10864){ if(EMOVIPMov$p124[i]==3 | EMOVIPMov$
    p124[i]==4) EMOVIPMov$p124[i]<-2}
for(i in 1:10864){ if(EMOVIPMov$p124[i]==5 | EMOVIPMov$
    p124[i]==6) EMOVIPMov$p124[i]<-3}
for(i in 1:10864){ if(EMOVIPMov$p124[i]==7 | EMOVIPMov$
    p124[i]==8) EMOVIPMov$p124[i]<-4}
for(i in 1:10864){ if(EMOVIPMov$p124[i]==9 | EMOVIPMov$
    p124[i]==10) EMOVIPMov$p124[i]<-5}

#Cambio de puntajes de clasificación de 1-5 del hogar
    actual

for(i in 1:10864){ if(EMOVIPMov$p169[i]==1 | EMOVIPMov$
    p169[i]==2) EMOVIPMov$p169[i]<-1}
for(i in 1:10864){ if(EMOVIPMov$p169[i]==3 | EMOVIPMov$
    p169[i]==4) EMOVIPMov$p169[i]<-2}
for(i in 1:10864){ if(EMOVIPMov$p169[i]==5 | EMOVIPMov$
    p169[i]==6) EMOVIPMov$p169[i]<-3}
for(i in 1:10864){ if(EMOVIPMov$p169[i]==7 | EMOVIPMov$
    p169[i]==8) EMOVIPMov$p169[i]<-4}
for(i in 1:10864){ if(EMOVIPMov$p169[i]==9 | EMOVIPMov$
    p169[i]==10) EMOVIPMov$p169[i]<-5}

#Asigna valor 1 si hubo movilidad y 0 lo contrario
for(i in 1:10864){ if(as.numeric(EMOVIPMov$p124[i])<as.
    numeric(EMOVIPMov$p169[i])) EMOVIPMov$movilidad[i]<-
    1 else EMOVIPMov$movilidad[i]<-0}

for(i in 1:10864){ if(EMOVIPMov$sex_ent[i]=="Mujer")
    EMOVIPMov$mov_M[i]<-EMOVIPMov$movilidad[i] else
    EMOVIPMov$mov_M[i]<-0}
for(i in 1:10864){ if(EMOVIPMov$sex_ent[i]=="Hombre")
    EMOVIPMov$mov_H[i]<-EMOVIPMov$movilidad[i] else
    EMOVIPMov$mov_H[i]<-0}
```

```
#Se inicializan otras variables
EMOVIPMov$wmov_M<-0
EMOVIPMov$wmov_H<-0
EMOVIPMov$u_k<-0
EMOVIPMov$wu_k<-0

for(i in 1:10864){EMOVIPMov$wmov_M[i]<- EMOVIPMov$
    ponderel_M[i]*EMOVIPMov$mov_M[i]}
for(i in 1:10864){EMOVIPMov$wmov_H[i]<- EMOVIPMov$
    ponderel_H[i]*EMOVIPMov$mov_H[i]}

#Obteniendo el promedio de las variables de interés
sumMM=sum(EMOVIPMov$wmov_M)
sumMH=sum(EMOVIPMov$wmov_H)
sumTM=sum(EMOVIPMov$ponderel_M)
sumTH=sum(EMOVIPMov$ponderel_H)
muM=sumMM/sumTM
muH=sumMH/sumTH

#Calculando el valor estimado de u_k
for(i in 1:10864){EMOVIPMov$u_k[i]<- ((EMOVIPMov$mov_M[
    i]-muM)/sumTM)-((EMOVIPMov$mov_H[i]-muH)/sumTH)}
for(i in 1:10864){EMOVIPMov$wu_k[i]<- EMOVIPMov$u_k[i]*
    EMOVIPMov$ponderel[i]}

##############################
#Estimación de la varianza#
##############################

#Se agrupan los valores por Municipios y Localidades no
    Urbanas
group_mov<-group_by(EMOVIPMov, strat, cve_mpo)
tab_mov<-summarise(group_mov, sum_ambos=sum(wu_k), n_a=
    n(), t_ta=(sum(wu_k)/n()))

#Se iniciliza una nueva variable
group_mov$wu_2<-0
for(j in 1:204){for(i in 1:10864){if(group_mov$strat[i
```

```
]==tab_mov$strat[j] & group_mov$cve_mpo[i]==tab_mov$
    cve_mpo[j]) group_mov$wu_2[i]<-((group_mov$wu_k[i]-
    tab_mov$t_ta[j])^2)}}

tab_mov<-summarise(group_mov, sum_ambos=sum(wu_k), n_a=
    n(), t_ta=(sum(wu_k)/n()), s_2=sum(wu_2))

#Se iniciliza una nueva variable
tab_mov$n_n<-0
for(j in 1:204) {tab_mov$n_n[j]<-((tab_mov$n_a[j]*tab_
    mov$s_2[j])/(tab_mov$n_a[j]-1))}

#Se calcula el valor de la estimación de varianza
var_mov=sqrt(sum(tab_mov$n_n))
vvarmov=sum(tab_mov$n_n)

#Se construye el intervalo de confianza con 95%
CI_IM=(muM-muH)-(1.64*var_mov)
CI_SM=(muM-muH)+(1.64*var_mov)
```

Bibliografía

Chambers, R. L. and Skinner, C. J. (2003). *Analysis of Survey Data*. Wiley, Inglaterra.

Cochran, W. G. (2000). *Técnicas de muestreo*. Compañía Editorial Continental, S.A., México.

Grajales, R. V., Campos V., R. M., and Huerta Wong, J. E. (2013). *Informe Movilidad Social en México 2013: Imagina tu futuro*. Centro de Estudios Espinosa Yglesias, México.

Hansen, M. H. and Hurwitz, W. N. (1943). *On the Theory of Sampling from Finite Populations*. The Annals of Mathematical Statistics 14. p. 333-362.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953). *Sample Survey Methods And Theory Volume I*. Wiley, E.U.A.

Heeringa, S. G., West, B. T., and Berglund, P. A. (2010). *Applied Survey Data Analysis*. Chapman and Hall/CRC, E.U.A.

Horvitz, D. and Thompson, D. J. (1952). *A Generalization of Sampling without Replacement from a Finite Universe*. Journal of the American Statistical Association 47. p. 663-685.

INEGI (2017a). Encuesta nacional de ingresos y gastos de los hogares 2016. enigh nueva serie. http://www.beta.inegi.org.mx/proyectos/enchogares/regulares/enigh/nc/2016/default.html.

INEGI (2017b). Relatoría y análisis del inegi. encuestas de ingresos y gastos de los hogares 2008-2016. http://www.beta.inegi.org.mx/contenidos/proyectos/investigacion/invenc/doc/relatoria_y_analisis_del_inegi.pdf.

Investigaciones Sociales, Políticas y de Opinión Pública, S. A. d. C. V. (2011). *Diseño de muestreo EMOVI-2011*. México.

Larson, H. J. (1969). *Introduction to Probability Theory and Statistical Inference*. Wiley, Nueva York.

Särndal, C. E., Swensson, B., and Wretman, J. H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, Nueva York.

Scheaffer, R. L., Mendenhall, I. W., Ott, R. L., and Gerow, K. G. (2011). *Elementary Survey Sampling*. Cengage Learning, E.U.A.

Solís, P. (2007). *Inequidad y movilidad social en Monterrey*. El Colegio de México, México.

Torche, F. (2009). *Sociological and Economic Approaches to the Intergenerational Transmission of Inequality in Latin America*. United Nations Development Programme (UNDP), Working Paper HD-09-2009.

von Mentz, B. (2003). *Movilidad social de sectores medios en México. Una retrospectiva histórica (siglos XVII al XX)*. Grupo Editorial Miguel Ángel Porrúa, México.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer-Verlag, Berlín.